Crop Price Prediction using

Collaborative Filtering, Classification, and Regression

Muhammad Abdullah, Moaz Shamsi

Punjab University College Of Information and Technology

**Abstract**

*Small and marginal farmers, who account for over 80% of Pakistan’s agricultural population, often sell their harvest at low, unfavorable prices before spoilage. These farmers often lack access to either cold storage or market forecasts. In particular, by having access to cold storage, farmers can store their produce for longer and thus have more flexibility as to when they should sell their harvest. Meanwhile, by having access to market forecasts, farmers can more easily identify which markets to sell at and when. While affordable cold storage solutions have become more widely available, there has been less work on produce price forecasting. A key challenge is that in many regions of Pakistan, predominantly in rural and remote areas, we have either very limited or no produce pricing data available from public online sources.*

*In this paper, we present a crop price forecasting system that pulls data from local markets, trains a model of prices using over a thousand markets, and displays interpretable price forecasts in a web application viewable from a mobile phone. Due to the pricing data being extremely sparse, our method first imputes missing entries using collaborative filtering to obtain a dense dataset. Using this imputed dense dataset, we then train a model to predict crop prices at different markets. In terms of interpretability, we display the most relevant historical pricing data that drive each forecasted price, where we take advantage of ensemble learning methods. We show how to construct heuristic price uncertainty intervals based on nearest neighbors. We validate forecast accuracy on data from a few local markets in Kasur.*

### **1 Introduction**

### Agriculture accounts for 18.9% of Pakistan’s GDP and is the principal source of income for roughly 42.3% of the pakistan population ( finance.gov.pk ). Small and marginal farmers, who have a field size of less than 2 hectares, make up over 80% of Pakistan’s agricultural population. Many of these farmers’ incomes are heavily tied to profits from selling their produce. To increase profits, these farmers could try to sell their produce at higher prices. However, the issue is that if they set the produce prices too high, then they run the risk of having a lot of leftover unsold produce. This unsold produce can then spoil and become unfit for consumption. To avoid having unsold produce, farmers often sell their produce at low, unfavorable prices. To make matters worse, smallholder farmers often devote a large portion of their fields to traditional food crops while using only a small fraction of their fields for cash crops. Two key problems are that small and marginal farmers often lack access to cold storage or to market information and forecasts.

In this paper, we develop a system for forecasting produce price trends while providing evidence for forecasts. Our system uses collaborative filter- ing to impute missing data and adaptive nearest neighbor methods to obtain interpretable forecasts (Section 3). For interpretability, we show which histor- ical pricing information (which markets and which calendar dates) drive any particular forecasted price change, and we also construct heuristic uncertainty intervals for forecasted prices (Section 3.4). We also present a web app that updates and displays price forecasts on a daily basis (Section 5)

2 Market Data Characteristics

We collect product pricing data from local govt agents named makerted commatte. Data is stored in hard documents, so we’ve typed data in sheets manually. Data contains a lot of missing entries, i.e A specific product at a specific market can have no recorded price or volume information across a large number of days. This missingness in data is caused by a market being closed, a produce not being sold on a particular day at a market, or the price or volume data for a produce at a market simply not being recorded even though the produce was indeed being sold and the market was open. We tried to impute missing entries using statistical techniques.

We remark that the prices we consider for both data sources are the modal price per produce for each day (when pricing data are actually available for that particular day). Meanwhile, we do not know what the rate of error is in verbal communication of prices or in data entry.

We’ll train our model at the local markets of Kasur, and predict results by using the govt retail rates on the given test dates.

3 Forecasting Method

In this section, we present our approach to forecasting produce prices. For sim- plicity, we train a different model per produce, so throughout this section, we assume that there is a single produce of interest. As input, we take pricing infor- mation (of the single produce of interest) from all Indian markets up to present time. As output, we forecast the direction of price change (up, down, or stay the same) per market at the next “time step”. What constitutes a “time step” is based on how time is quantized. We present results later where one time step corresponds to 1 day, 1 week, 2 weeks, or 4 weeks. Our forecasting approach can easily be modified to predict exact prices at the next time step (regression) rather than just the direction of price changes (classification). However, we

emphasize that predicting exact prices is a much harder problem than predict- ing the direction of price changes. For clarity of exposition, we introduce our method in the classification context. We defer discussing the regression context to Section

Our forecasting method consists of three steps. First, we impute missing data using a standard collaborative filtering method; at this point, we use the finest level of granularity for time in the raw input data. In the second step, we quantize time. Lastly, using the dense imputed data that have been quantized in time, we train a model using an adaptive nearest neighbor method such as random forests or boosting methods that use decision trees as base predictors. A pictorial overview of these three steps is shown in Figure 3. We explain these three steps in detail in the next three subsections.

We remark that for the third step, we intentionally use adaptive nearest neighbor methods because they can readily provide “evidence” for forecasts (Chen and Shah, 2018, Section 7.1). In our problem context, this means that for any predicted price change direction, an adaptive nearest neighbor method can provide historical dates and prices that are directly used in making the prediction; these historical dates and prices can be supplied to the farmer as forecast evidence and, as we discuss in Section 3.4, can also be used to construct a heuristic “uncertainty interval”. This uncertainty interval can be thought of as a range of plausible prices associated with a forecasted price change direction. Instead of adaptive nearest neighbor methods, other machine learning classifiers could be used instead although providing forecast evidence and some notion of a price uncertainty interval may be difficult depending on the method used.